Rate-Information-Optimal Gaussian Channel Output Compression

Andreas Winkelbauer and Gerald Matz

Institute of Telecommunications, Vienna University of Technology
Gusshausstrasse 25/389, 1040 Vienna, Austria
email: {andreas.winkelbauer, gerald.matz}@nt.tuwien.ac.at

Abstract—We study the maximum rate achievable over a Gaussian channel with Gaussian input under channel output compression. This problem is relevant to receive signal quantization in practical communication systems. We use the Gaussian information bottleneck to provide closed-form expressions for the information-rate function and the rate-information function, which quantify the optimal trade-off between the compression rate and the corresponding end-to-end mutual information. We furthermore show that mean-square error optimal compression of the channel output achieves the optimal trade-off, thereby greatly facilitating the design of channel output quantizers.

Index Terms—quantization, source coding, rate distortion theory, information bottleneck, data compression

I. INTRODUCTION

Quantization is of fundamental importance for modern communication systems. In fact, every digital receiver in a communication system has to employ some form of quantization, e.g., analog-to-digital conversion. Quantization is well understood in the lossy source coding setting. However, a lossy source coding perspective is generally not appropriate in the context of communication systems where we are interested in maximizing the data rate rather than in accurately representing the received signal.

In this paper, we investigate the fundamental limits for data transmission over a Gaussian channel with Gaussian input under channel output compression. In particular, we find the relation between the compression rate and the resulting mutual information. We term this relation the *rate-information trade-off*. Furthermore, we discuss information-theoretically optimal quantization. Specifically, our contributions are as follows.

- We determine the optimal rate-information trade-off in terms of the *Gaussian information bottleneck* (GIB) [1].
- We define the *information-rate function* and the *rate-information function* and discuss their properties.
- We show that *rate distortion* (RD) theory [2] yields the optimal rate-information trade-off if the squared error is chosen as distortion metric.
- We discuss the design and the properties of quantizers that are optimal in an information-theoretic sense.
- We briefly discuss generalizations of our results to the complex-valued case and to Gaussian vector channels.

The remainder of this paper is organized as follows. Section II summarizes the required background. In Section III, we present the optimal rate-information trade-off and its properties. The equivalence of GIB and RD theory for Gaussian channels and squared-error distortion is established in Section IV. Section V considers the design and the properties of optimized channel output quantizers. Generalizations and conclusions are provided in Sections VI and VII, respectively.

Notation: We use boldface letters for column vectors and upright sans-serif letters for random variables. The expectation operator is denoted by $\mathbb{E}$ and we follow the notation of [3] for entropy $H(\cdot)$, differential entropy $h(\cdot)$, and mutual information $I(\cdot;\cdot)$. The identity matrix is denoted by $\boldsymbol{I}$ and $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{C})$ denotes a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$. Furthermore, $[x]^+ \triangleq \max\{0, x\}$ and $\log^+ x \triangleq [\log x]^+$. All logarithms are to base 2.

II. BACKGROUND AND DEFINITIONS

A. Gaussian Information Bottleneck

The *information bottleneck method* [4] and its variant, the GIB, so far have received limited attention outside of the machine learning community. We therefore give a brief overview of the GIB.

Let $\mathbf{x}$ and $\mathbf{y}$ be jointly Gaussian zero-mean random vectors with full rank covariance matrices. We denote the length of $\mathbf{y}$ by n. Furthermore, let $\mathbf{z}$ be a compressed representation of $\mathbf{y}$ that is characterized by the conditional distribution $p(\mathbf{z}|\mathbf{y})$. The compression rate equals $I(\mathbf{y}; \mathbf{z})$. It follows that $\mathbf{x} - \mathbf{y} - \mathbf{z}$ forms a Markov chain. The GIB addresses the following variational problem[1]:

$$\min_{p(\mathbf{z}|\mathbf{y})} I(\mathbf{y}; \mathbf{z}) - \beta I(\mathbf{x}; \mathbf{z}). \tag{1}$$

In the context of the information bottleneck method, $\mathbf{x}$ is called the *relevance variable* and $I(\mathbf{x}; \mathbf{z})$ is termed *relevant information*. The trade-off between compression rate and relevant information is determined by the positive parameter β. In [6] it has been shown that the optimal $\mathbf{z}$ is jointly Gaussian with $\mathbf{y}$ and can therefore be written as

$$\mathbf{z} = \boldsymbol{A}\mathbf{y} + \boldsymbol{\xi}, \tag{2}$$

---

[1]A related problem (formulated in terms of conditional entropy) involving a Markov chain of three discrete random variables is studied in [5].
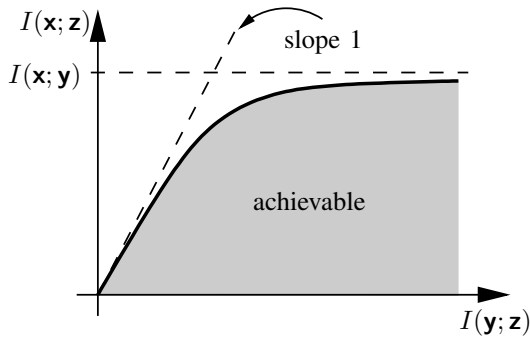
Figure 1: Illustration of the trade-off between compression rate $I(\mathbf{y};\mathbf{z})$ and relevant information $I(\mathbf{x};\mathbf{z})$.

where $\boldsymbol{A}$ is an $n \times n$ matrix and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C_\xi})$ is independent of $\mathbf{y}$. Hence, we can rewrite the problem in (1) using (2) as

$$\min_{\boldsymbol{A}, \boldsymbol{C_\xi}} I(\mathbf{y}; \boldsymbol{A}\mathbf{y} + \boldsymbol{\xi}) - \beta I(\mathbf{x}; \boldsymbol{A}\mathbf{y} + \boldsymbol{\xi}). \tag{3}$$

A solution of (3) is given by [1, Theorem 3.1]

$$\boldsymbol{A} = \operatorname{diag}\{\alpha_k\}_{k=1}^n \boldsymbol{V}^{\mathrm{T}} \qquad \text{and} \qquad \boldsymbol{C_\xi} = \boldsymbol{I}. \tag{4}$$

Here, $\boldsymbol{V} = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_n]$ is the matrix of left eigenvectors of $\boldsymbol{C}_{\mathbf{y}|\mathbf{x}} \boldsymbol{C}_{\mathbf{y}}^{-1} = \mathbb{E}\{\mathbf{y}\mathbf{y}^{\mathrm{T}}|\mathbf{x}\} (\mathbb{E}\{\mathbf{y}\mathbf{y}^{\mathrm{T}}\})^{-1}$ and

$$\alpha_k = \sqrt{\frac{[\beta(1-\lambda_k)-1]^+}{\lambda_k \boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{C}_{\mathbf{y}} \boldsymbol{v}_k}}, \qquad k = 1, \ldots, n, \tag{5}$$

with λ_k the eigenvalue corresponding to $\boldsymbol{v}_k$. Using (4) and (5), the rate-information trade-off can be expressed as follows:

$$I(\mathbf{x};\mathbf{z}) = I(\mathbf{y};\mathbf{z}) - \frac{1}{2} \sum_{k=1}^n \log^+ \beta(1-\lambda_k). \tag{6}$$

By the data processing inequality, (6) is bounded as follows:

$$I(\mathbf{x};\mathbf{z}) \le \min\{I(\mathbf{y};\mathbf{z}), I(\mathbf{x};\mathbf{y})\}. \tag{7}$$

Fig. 1 illustrates (6) (solid line) and (7) (dashed lines); both $I(\mathbf{x};\mathbf{z})$ and $I(\mathbf{y};\mathbf{z})$ increase as β increases. The shaded region in Fig. 1 corresponds to the achievable rate-information pairs (see [7] for details).

B. Information-Rate Function and Rate-Information Function

To formalize the trade-off between relevant information and compression rate, we next define the *information-rate function* $I(R)$ and the *rate-information function* $R(I)$.

Definition 1. *Let* $\mathbf{x}-\mathbf{y}-\mathbf{z}$ *be a Markov chain. The information-rate function* $I: \mathbb{R}_+ \to [0, I(\mathbf{x};\mathbf{y})]$ *is defined by*

$$I(R) \triangleq \max_{p(\mathbf{z}|\mathbf{y})} I(\mathbf{x};\mathbf{z}) \quad \text{subject to } I(\mathbf{y};\mathbf{z}) \le R. \tag{8}$$

The rate-information function $R: [0, I(\mathbf{x};\mathbf{y})] \to \mathbb{R}_+$ *is defined by*

$$R(I) \triangleq \min_{p(\mathbf{z}|\mathbf{y})} I(\mathbf{y};\mathbf{z}) \quad \text{subject to } I(\mathbf{x};\mathbf{z}) \ge I. \tag{9}$$

$I(R)$ allows us to quantify the maximum of the relevant information that can be preserved when the compression rate

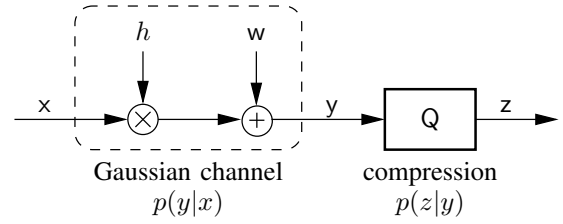is at most R. Conversely, $R(I)$ quantifies the minimum compression rate required when the retained relevant information must be at least I. We note that the "distance" between $\mathbf{y}$ and $\mathbf{z}$ is immaterial for $I(R)$ and $R(I)$. This is in contrast to the structurally similar definitions of the rate-distortion and distortion-rate functions.

III. THE RATE-INFORMATION TRADE-OFF

We next derive $I(R)$ and $R(I)$ in closed-form for the real-valued scalar Gaussian channel shown in Fig. 2 (see Sec. VI for generalizations). We can express any zero-mean jointly Gaussian random variables x, y as

$$\mathsf{y} = h\mathsf{x} + \mathsf{w}, \tag{10}$$

where $h \in \mathbb{R}$ and $\mathsf{w} \sim \mathcal{N}(0, \sigma^2)$ is independent of x. Setting $\mathsf{x} \sim \mathcal{N}(0, P)$ yields $\mathsf{y} \sim \mathcal{N}(0, h^2 P + \sigma^2)$. The compressed representation of y is denoted $\mathsf{z} = \mathsf{Q}(\mathsf{y})$. By Markovity of $\mathsf{x} - \mathsf{y} - \mathsf{z}$ we have

$$p(z|x) = \int_{-\infty}^{\infty} p(z|y) p(y|x) \mathrm{d}y, \tag{11}$$

where $p(y|x)$ is the transition pdf of the Gaussian channel and $p(z|y)$ describes the compression mapping Q.

The capacity of the Gaussian channel $p(y|x)$ with average power constraint P and no channel output compression equals [3, Sec. 10.1]

$$C(\rho) \triangleq \frac{1}{2} \log(1 + \rho) \tag{12}$$

with the *signal-to-noise ratio* (SNR)

$$\rho \triangleq \frac{h^2 P}{\sigma^2}. \tag{13}$$

The following theorem states a closed-form expression for the information-rate function and discusses its properties.

Theorem 2. *The information-rate function of a Gaussian channel with SNR ρ is given by*

$$I(R) = R - \frac{1}{2} \log \frac{2^{2R} + \rho}{1 + \rho} \tag{14}$$

$$= C(\rho) - C(2^{-2R}\rho). \tag{15}$$

$I(R)$ *has the following properties:*
1) $I(R)$ *is strictly concave on* $\mathbb{R}_+$.
2) $I(R)$ *is strictly increasing in* R.
3) $I(R) \le \min\{R, C(\rho)\}$.
4) $I(0) = 0$ *and* $\lim_{R \to \infty} I(R) = C(\rho)$.
5) $\frac{\mathrm{d}I(R)}{\mathrm{d}R} = (1 + 2^{2R}\rho^{-1})^{-1} \le \frac{\mathrm{d}I(R)}{\mathrm{d}R}\big|_{R=0} = (1 + \rho^{-1})^{-1}$.

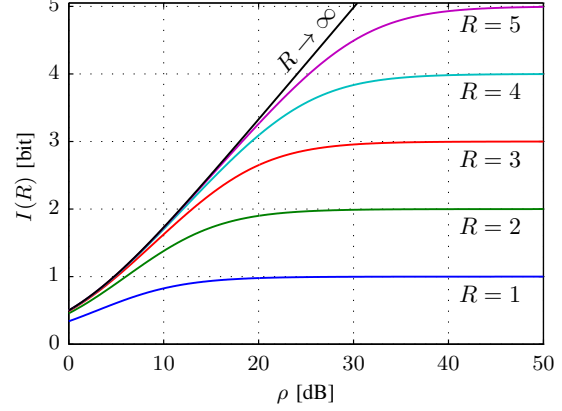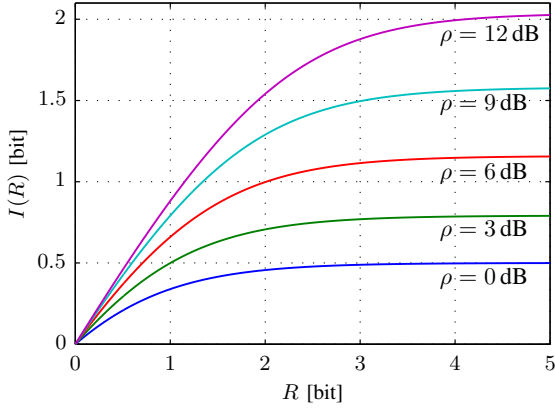Figure 2: Gaussian channel with channel output compression.

Figure 3: (a) $I(R)$ vs. R for various SNRs ρ; (b) $I(R)$ vs. ρ for various rates R.

Proof: Due to the GIB, the optimal z equals

$$z = ay + \xi = ahx + aw + \xi \qquad (16)$$

with $\xi \sim \mathcal{N}(0,1)$ independent of y and

$$a = \sqrt{\frac{[\beta(1 + \rho^{-1})^{-1} - 1]^+}{\sigma^2}}. \qquad (17)$$

It follows from (16) that $z|x \sim \mathcal{N}(ahx, a^2\sigma^2 + 1)$, i.e., the overall channel $p(z|x)$ is again Gaussian. With $z \sim \mathcal{N}(0, a^2(h^2P + \sigma^2) + 1)$ and $z|y \sim \mathcal{N}(ay, 1)$, we can express the compression rate as

$$R \triangleq I(y; z) = h(z) - h(z|y) \qquad (18)$$

$$= \frac{1}{2} \log \rho(\beta - 1), \qquad (19)$$

where we have used (17) and the fact that the differential entropy of a Gaussian random variable with variance σ^2 is $\frac{1}{2}\log(2\pi e\sigma^2)$. From (19) we can express β in terms of R as

$$\beta(R) = 1 + \frac{2^{2R}}{\rho}. \qquad (20)$$

The relevant information can be calculated as

$$I \triangleq I(x; z) = h(z) - h(z|x) \qquad (21)$$

$$= R - \frac{1}{2} \log \beta(1 + \rho^{-1})^{-1}. \qquad (22)$$

Finally, using (20) together with (22) yields the information-rate function. The properties of $I(R)$ are easily verified from (14) and (15). ∎

From (14) we conclude that $I(R) \approx R$ for small R. Similarly, (15) implies $I(R) \approx C(\rho)$ for large R. We call $R < C(\rho)$ the *compression-limited* regime ($I(R) < R$) and $R > C(\rho)$ the *noise-limited* regime ($I(R) < C(\rho)$). In Fig. 3a we plot the information-rate function for different values of ρ. Here, the curves saturate at the respective channel capacity as R becomes large. In Fig. 3b we plot $I(R)$ vs. ρ for different values of R. In this case, the curves saturate at R as the SNR becomes large. The following corollaries are due to the fact that the overall channel $p(z|x)$ is Gaussian.

Corollary 3. *We can rewrite $I(R)$ in (14) as follows:*

$$I(R) = C(\hat{\rho}), \qquad (23)$$

where the SNR $\hat{\rho}$ of the overall channel $p(z|x)$ is

$$\hat{\rho} = \rho \frac{1 - 2^{-2R}}{1 + 2^{-2R}\rho} \leq \rho. \qquad (24)$$

Corollary 4. *The channel output compression can equivalently be modeled by an additive zero-mean Gaussian noise term with variance*

$$\sigma_q^2 = \sigma^2 \frac{1 + \rho}{2^{2R} - 1}, \qquad (25)$$

yielding

$$\hat{\rho} = \frac{\rho}{1 + \frac{\sigma_q^2}{\sigma^2}}. \qquad (26)$$

In the next theorem, we give a closed-form expression for the rate-information function and discuss its properties.

Theorem 5. *The rate-information function of a Gaussian channel with SNR ρ is given by*

$$R(I) = \frac{1}{2} \log \frac{\rho}{2^{-2I}(1 + \rho) - 1}. \qquad (27)$$

$R(I)$ has the following properties:

1) $R(I)$ is strictly convex on $[0, C(\rho)]$.
2) $R(I)$ is strictly increasing in I.
3) $R(I) \geq I$.
4) $R(0) = 0$ and $\lim_{I \to C(\rho)} R(I) = \infty$.
5) $\frac{dR(I)}{dI} = (1 + \rho)/(1 + \rho - 2^{2I}) \geq \frac{dR(I)}{dI}\big|_{I=0} = 1 + \rho^{-1}$.

Proof: (27) follows directly from (14). The properties of $R(I)$ are easily verified from (27). ∎

Corollary 6. *The rate-information function (27) is the inverse of the information-rate function (14), i.e., for $\tilde{I} \in [0, C(\rho)]$ and $\tilde{R} \in \mathbb{R}_+$ we have*

$$I(R(\tilde{I})) = \tilde{I} \quad \text{and} \quad R(I(\tilde{R})) = \tilde{R}. \tag{28}$$

Thus, the derivatives of $I(R)$ and $R(I)$ are related as

$$I'(\tilde{R}) = \frac{1}{R'(I(\tilde{R}))} \quad \text{and} \quad R'(\tilde{I}) = \frac{1}{I'(R(\tilde{I}))}. \tag{29}$$

Corollary 7. *Let $\varepsilon > 0$. The minimum compression rate R_ε required to achieve $I(R) \geq (1 - \varepsilon)C(\rho)$ equals*

$$R_\varepsilon = \frac{1}{2} \log \frac{\rho}{(1 + \rho)^\varepsilon - 1}. \tag{30}$$

IV. EQUIVALENCE TO MSE-OPTIMAL COMPRESSION

We next show that optimal compression of the channel output in the RD sense also yields the optimal rate-information trade-off if squared-error distortion is used. However, we emphasize that this result holds only for real-valued scalar Gaussian channels (see Sec. VI for generalizations). In what follows we denote the rate-information trade-off achievable by RD-optimal compression as $I^{\mathrm{RD}}(R)$ and $R^{\mathrm{RD}}(I)$.

Theorem 8. *The rate-information trade-off achievable by RD-optimal compression of the output of a Gaussian channel with SNR ρ is given by*

$$I^{\mathrm{RD}}(R) = \frac{1}{2} \log \frac{1 + \rho}{1 + 2^{-2R}\rho} \tag{31}$$

if the square-error is used as distortion measure. $I^{\mathrm{RD}}(R)$ equals the information-rate function, i.e., $I^{\mathrm{RD}}(R) = I(R)$ and therefore $R^{\mathrm{RD}}(I) = R(I)$. Hence, RD-optimal compression of the channel output using squared-error distortion is optimal also in terms of the rate-information trade-off.

Proof: RD-optimal compression of $\mathsf{y} \sim \mathcal{N}(0, h^2 P + \sigma^2)$ with rate $R = I(\mathsf{y}; \mathsf{z})$ yields $\mathsf{y}|\mathsf{z} \sim \mathcal{N}(z, D(R))$ with the *mean-square error* (MSE) distortion [3, Sec. 13.3]

$$D(R) = 2^{-2R}(h^2 P + \sigma^2). \tag{32}$$

We first find $p(x|z)$ to derive $I^{\mathrm{RD}}(R)$. We have

$$p(x|z) = \int_{-\infty}^{\infty} p(x|y)p(y|z)\mathrm{d}y \tag{33}$$

$$= p(x) \int_{-\infty}^{\infty} \frac{p(y|x)}{p(y)} p(y|z)\mathrm{d}y, \tag{34}$$

where (33) is due to Markovity of $\mathsf{x}-\mathsf{y}-\mathsf{z}$. Using $\mathsf{x} \sim \mathcal{N}(0, P)$ and $\mathsf{y}|\mathsf{x} \sim \mathcal{N}(hx, \sigma^2)$, the integral in (34) can be written as

$$\int_{-\infty}^{\infty} \frac{p(y|x)}{p(y)} p(y|z)\mathrm{d}y = \sqrt{\frac{2^{2R}}{2\pi\sigma^2}} \exp\left(-\frac{h^2 x^2}{2\sigma^2} - \frac{z^2}{2D(R)}\right)$$
$$\times \int_{-\infty}^{\infty} \exp(Ay^2 + By)\mathrm{d}y, \tag{35}$$

where

$$A = \frac{1}{2(h^2 P + \sigma^2)} - \frac{1}{2\sigma^2} - \frac{1}{2D(R)}, \tag{36}$$

$$B = \frac{hx}{\sigma^2} + \frac{z}{D(R)}. \tag{37}$$

Completing the square in the integral on the right-hand side of (35) yields

$$\int_{-\infty}^{\infty} \exp(Ay^2 + By)\mathrm{d}y = \exp\left(-\frac{B^2}{4A}\right)\sqrt{-\frac{\pi}{A}}. \tag{38}$$

Using (35)-(38) in (34), rearranging terms and simplifying expressions finally yields $\mathsf{x}|\mathsf{z} \sim \mathcal{N}(\mu, \varsigma^2)$, where the mean and the variance are respectively given by

$$\mu = \frac{z}{h} \frac{\rho}{1 + \rho}, \tag{39}$$

$$\varsigma^2 = P \frac{1 + 2^{-2R}\rho}{1 + \rho}. \tag{40}$$

We thus obtain $I^{\mathrm{RD}}(R) = I(\mathsf{x}; \mathsf{z})$ as

$$I(\mathsf{x}; \mathsf{z}) = h(\mathsf{x}) - h(\mathsf{x}|\mathsf{z}) \tag{41}$$

$$= \frac{1}{2} \log \frac{1 + \rho}{1 + 2^{-2R}\rho}, \tag{42}$$

which is equal to $I(R)$ in (14). ∎

Theorem 8 is surprising because minimizing the MSE $\mathbb{E}\{(\mathsf{y}-\mathsf{z})^2\}$ in a Markov chain $\mathsf{x}-\mathsf{y}-\mathsf{z}$ subject to $I(\mathsf{y}; \mathsf{z}) \leq R$ will generally not maximize the mutual information $I(\mathsf{x}; \mathsf{z})$ with respect to $p(z|y)$. However, in the case of real-valued, scalar, and jointly Gaussian x, y this is indeed the case. Other examples where MSE-optimal processing is also information-theoretically optimal can be found in [8] and [9].

V. OPTIMAL DESIGN OF CHANNEL OUTPUT QUANTIZERS

In this section, we study deterministic channel output (vector) quantizers which are optimal in an information-theoretic sense. Due to Theorem 8, we know that minimizing the MSE distortion at a given rate maximizes $I(\mathsf{x}; \mathsf{z})$. We note that

$$I(R) \geq I(D(R)) = \frac{1}{2} \log \frac{1 + \rho}{1 + \frac{D(R)}{h^2 P + \sigma^2}\rho}, \tag{43}$$

where

$$I(R) = \max_{D(R)} I(D(R)). \tag{44}$$

We consider channel output quantizers with finite blocklength (i.e., vector quantizer dimension). We propose to design the quantizers such that the lower bound in (43) is maximized, which is achieved by minimizing the MSE $\mathbb{E}\{(\mathsf{y} - \mathsf{z})^2\}$. We note that this strategy is asymptotically optimal due to (44).

An important property of MSE-optimal quantizers is that their quantization regions are disjoint convex sets [10, Theorem 1]. We note that convex quantization regions are not always optimal in quantizer design for communication problems (see, e.g., [11] for a counterexample). The existence of an MSE-optimal partition of the input space also implies that randomized quantization cannot improve upon deterministic
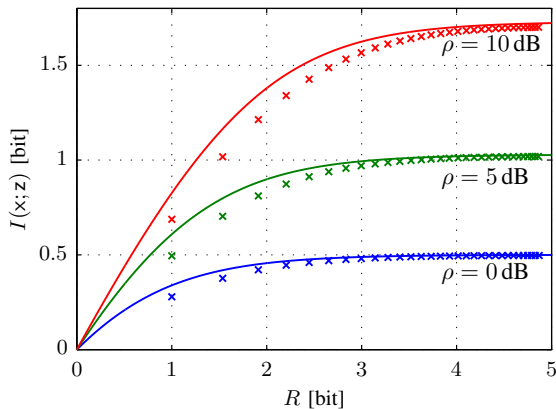
Figure 4: Solid lines correspond to $I(R)$ and '×' markers correspond to quantizers with 2 to 32 quantization levels.

quantization. This is because any randomized quantizer can be realized by using a set of (possibly suboptimal) deterministic quantizers in a time-sharing manner.

The fact that MSE-optimal quantizers with convex quantization regions are sufficient is very convenient for the quantizer design. Specifically, optimal quantizers can be designed using the Lloyd algorithm [12] and the LBG algorithm [13]. Fig. 4 shows how close we can get to $I(R)$ using scalar quantizers for $\rho \in \{0\,\text{dB}, 5\,\text{dB}, 10\,\text{dB}\}$. We have used the Lloyd algorithm to design quantizers with 2 to 32 quantization levels (here, $R = I(\mathsf{y}; \mathsf{z}) = H(\mathsf{z})$). We observe that the gap to $I(R)$ decreases as R increases and at fixed rate the gap to $I(R)$ grows with increasing SNR.

VI. GENERALIZATIONS

A. Complex-Valued Gaussian Channels

The results in this paper straightforwardly generalize to the case where x, y are scalar *circularly symmetric* complex-valued jointly Gaussian random variables. All equations have the same structure with the only significant differences being that the pre-log factor of $\frac{1}{2}$ vanishes and $2^{\pm 2R}$, $2^{\pm 2I}$ is replaced by $2^{\pm R}$, $2^{\pm I}$. In particular, the information-rate function and the rate-information function read

$$I(R) = R - \log \frac{2^R + \rho}{1 + \rho}, \qquad (45)$$

$$R(I) = \log \frac{\rho}{2^{-I}(1 + \rho) - 1}, \qquad (46)$$

where ρ is the SNR of the real part and the imaginary part.

We note that our results do not hold if x, y are not circularly symmetric. In this case we moreover have $I^{\text{RD}}(R) < I(R)$ (see [14] for details).

B. Gaussian Vector Channels

The extension to Gaussian vector channels is not treated here due to lack of space. We derive the rate-information trade-off for Gaussian vector channels in [14]. It turns out that in this case MSE-optimal processing does not yield the optimal

rate-information trade-off. We have $I(R) - I^{\text{RD}}(R) \geq 0$ and the difference essentially depends on the eigenvalue spread of the channel. However, MSE-optimal compression preceded by a particular linear filter is sufficient to achieve the optimal rate-information trade-off [15].

VII. CONCLUSIONS

We have studied the maximum achievable rate of scalar Gaussian channels with Gaussian input under channel output compression. We have used the GIB to derive the optimal rate-information trade-off in terms of the information-rate function and the rate-information function. It turns out that optimal channel output compression can be modeled by an additive Gaussian noise term. We have shown that here RD-optimal channel output compression using squared-error distortion also achieves the optimal rate-information trade-off. This fact greatly simplifies the design of optimal channel output quantizers, since it is sufficient to consider MSE-optimal quantizers. Finally, we have briefly discussed generalizations of our results that will be covered in more detail elsewhere.

REFERENCES

[1] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables," *Journal of Machine Learning Research*, vol. 6, pp. 165–188, Jan. 2005.

[2] T. Berger, *Rate Distortion Theory*. Englewood Cliffs (NJ): Prentice Hall, 1971.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[4] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Allerton Conf. on Communication, Control, and Computing*, Sept. 1999, pp. 368–377.

[5] H. Witsenhausen and A. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 493–501, Sept. 1975.

[6] A. Globerson and N. Tishby, "On the optimality of the Gaussian information bottleneck curve," The Hebrew University of Jerusalem, Tech. Rep., Feb. 2004.

[7] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Learning Theory and Kernel Machines*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2003, vol. 2777, pp. 595–609.

[8] T. Guess and M. K. Varanasi, "An information-theoretic derivation of the MMSE decision-feedback equalizer," in *Proc. 36th Allerton Conf. on Communication, Control, and Computing*, Sept. 1998.

[9] G. D. Forney, "On the role of MMSE estimation in approaching the information-theoretic limits of linear gaussian channels: Shannon meets Wiener," in *Proc. 41st Allerton Conf. on Communication, Control, and Computing*, Oct. 2003, pp. 430–439.

[10] D. Burshtein, V. D. Pietra, D. Kanevsky, and A. Nadas, "Minimum impurity partitions," *The Annals of Statistics*, vol. 20, no. 3, pp. 1637–1646, Sept. 1992.

[11] G. Zeitler, "Low-precision analog-to-digital conversion and mutual information in channels with memory," in *Proc. 48th Allerton Conf. on Communication, Control, and Computing*, Sept. 2010, pp. 745–752.

[12] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, March 1982.

[13] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Comm.*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[14] A. Winkelbauer, S. Farthofer, and G. Matz, "The rate-information trade-off for Gaussian vector channels," submitted to *IEEE Int. Symp. Information Theory*, 2014.

[15] M. Meidlinger, A. Winkelbauer, and G. Matz, "On the relation between the Gaussian information bottleneck and MSE-optimal rate-distortion quantization," submitted to *IEEE Workshop on Statistical Signal Processing*, 2014.